4-18-2013

Language Encodes Social Network Information

Sterling Chelsea Hutchinson

LANGUAGE STATISTICS ENCODES SOCIAL NETWORK INFORMATION

by

Sterling Chelsea Hutchinson

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Psychology

The University of Memphis

May 2013

Abstract

Hutchinson, Sterling Chelsea. MS. The University of Memphis. May 2013. Language Encodes Social Network Information. Major Professor: Max Louwerse, Ph.D.

Knowledge regarding social information is commonly thought to be derived from sources such as interviews and formal relationships. Consequently, social networks can be generated from this information. Recent work has demonstrated that language statistics can explain findings often thought to primarily be explained by external factors. Three studies explored whether language implicitly comprises information that allows for extracting social networks, by testing the hypothesis that individuals who are socially related together are linguistically discussed together, as well as the hypothesis that individuals who are socially related more are linguistically discussed more. Three computational studies were conducted testing the extent to which social networks could be extracted from fiction novels. Semantic relationships revealed that MDS solutions correlated with the actual social network of characters. A human study in which participants estimated social relationships of characters matched the results obtained computationally. The results demonstrated that linguistic information encodes social relationship information.

TABLE OF CONTENTS

Page

1. Introduction	1
2. Study 1: Computational Study with Few Characters	6
Manually Generated Social Network Maps	7
First-order co-occurrences	8
Higher-order co-occurrences	12
Number of Relationships	16
Discussion	16
3. Study 2: Computational Study with Many Characters	17
Manually Generated VUE Social Network Map	17
First-order co-occurrences	18
Higher-order co-occurrences	19
Number of Relationships	24
Discussion	24
4. Study 3a: Computational Study with Moderate Number of Characters	24
Manually Generated Social Network Maps	25
First-order co-occurrences	25
Higher-order co-occurrences	28
Number of Relationships	31
5. Study 3b: Human Study with Moderate Number of Characters	31
Method	32
First-order co-occurrences	34
Higher-order co-occurrences	37
Number of Relationships	40
Discussion	40
6. General Discussion	40
References	44
Appendices	
A. Questionnaire designed to test knowledge of the Harry Potter books	48
B. IRB Review and Approval	50

LIST OF FIGURES

Figure | Page

Figure 1. Correlation between bigram values and Muckety map values in *Twilight* for the first dimension representing character prominence 10

Figure 2. Correlation between bigram values and Muckety map values in *Twilight* for the second dimension representing friends and adversaries 11

Figure 3. Correlation between LSA values and Muckety map values in *Twilight* for the first dimension representing character prominence 13

Figure 4. Correlation between LSA values and Muckety map values in *Twilight* for the second dimension representing friends and adversaries 14

Figure 5. Correlation between bigram values and VUE complex map values in *A Song of Ice and Fire* for the first dimension representing character prominence 20

Figure 6. Correlation between bigram values and VUE complex map values in *A Song of Ice and Fire* for the second dimension representing friends and adversaries 21

Figure 7. Correlation between LSA values and VUE complex map values in *A Song of Ice and Fire* for the first dimension representing character prominence 22

Figure 8. Correlation between LSA values and VUE complex map values in *A Song of Ice and Fire* for the second dimension representing friends and adversaries 23

Figure 9. Correlation between bigram values and Muckety values in *Harry Potter* for the first dimension representing character prominence 26

Figure 10. Correlation between bigram values and Muckety values in *Harry Potter* for the second dimension representing friends and adversaries 27

Figure 11. Correlation between LSA values and Muckety values in *Harry Potter* for the first dimension representing character prominence 29

Figure 12. Correlation between LSA values and Muckety values in *Harry Potter* for the second dimension representing friends and adversaries 30

Figure 13. Correlation between bigram values and human map values in *Harry Potter* for the first dimension representing character prominence 35

Figure 14. Correlation between bigram values and human map values in *Harry Potter* for the second dimension representing friends and adversaries 36

Figure 15. Correlation between LSA values and human map values in *Harry Potter* for the second dimension representing friends and adversaries 38

Figure 16. Correlation between LSA values and human map values in *Harry Potter* for the second dimension representing friends and adversaries 39

1. Introduction

Fewer than 100 friends on Facebook? You must be socially isolated! Happily our social environment is not determined by friendships in social media, but rather by the friends we know and care about, our family, and our colleagues. What is the nature of these social relations and how can these relations be determined?

To answer this question, social scientists argue that social relationships can be explained by three dimensions (Fischer, 1982). First, social relationships are formalized as socially recognized roles, such as teacher/student, employer/employee, or father/son. Second, relationships can be sentimental, for instance when people feel close to one another. Third, a relation can be defined in terms of functional interactions and exchanges. This formal, sentimental, and interactive nature of a social relationship is determined by behaviors, emotions, and environments, among other factors that impact the trajectory of relationships. For instance, environments tend to weigh heavily in terms of whether or not two individuals are likely to build a relationship together, with proximity having long been established as a strong predictor of relationships (Festinger, Schachter, & Black, 1950). In addition, ties between locations (e.g., commonly trekked routes) also impact social interaction (Takhteyev, Gruzd, & Wellman, 2011). Not only does increased physical proximity lead to increased likelihood of interpersonal relationships (Ebbesen, Kjos, & Konecni, 1976), both close physical and temporal proximity are actually excellent indicators of social ties between individuals (Crandall et al., 2010).

Similarly, familiarity fosters attraction between individuals (Reis, Manianci, Caprariello, Eastwick, & Finkel, 2011; Zajonc, 1968; 2001). Those who share interests, attitudes, and characteristics are more likely to develop friendships. In fact, any similarity between two individuals promotes the formation of a relationship between them (Byrne, 1971) with important matters (e.g., religious views, political attitudes) being given more weight (Touhey, 1972). Emotions too impact relationships. When two individuals first encounter one another, a future friendship becomes more likely if the interaction is positive, whereas a friendship is not apt to blossom if the interaction is negative (Farina, Wheeler, & Mehta, 1991). Even physical features, like smell or appearance influence the social relationships we form (Li, Moallem, Paller, & Gottfried, 2007).

Once these relationships have been established, even more factors come into play. For instance, Granovetter (1973) found that relationship strength was impacted by duration, intimacy, emotional intensity, and reciprocation. Indeed, researchers have found that factors like individual differences, attachment styles, and equity impact the duration of social relationships (Feeney & Noller, 1992; Hatfield, Traupmann, & Walster, 1978). Sprecher and Henrick (2004) found that reported levels of self-disclosure were also significantly related to satisfaction of established relationships.

With the many factors above impacting social relationships, how are these networks plotted? Social networks are structures that map relationships between individuals. They are complex systems that can be used to examine, predict, and measure various features embedded within a network (see Newman, 2003 for an overview). Nodes represent specific individuals with edges connecting those individuals and representing relational information. There are several ways these social networks are produced. Social networks are often generated manually whereby individuals are linked to others if they are friends, colleagues, family members, etc. Individuals are able to generate their own egocentric social networks representing those other individuals with whom they share a relationship. Of course, the individual generating the network will do so based on the existence and strength of relationships that were generated by, and subject to, the factors enumerated above (Scott, 1988).

In contrast to a self-generated network, social scientists often rely on interviews by asking individuals to list their friends, family, and colleagues, thereby manually generating a network (Fischer, 1982). Alternatively, instead of interviewing individuals to obtain a social network, relationships can be directly measured by actual physical and temporal distances between individuals. Like networks plotted through information garnered from interviews and from direct physical proximity, information from self reports about direct interactions between individuals can also help establish a network of social relationships for any given individual. But how can such networks be represented when participants cannot be interviewed as in Fischer's study, or when participants otherwise do not voluntarily release personal information as in self-generated networks and self reported measures? Can such networks be generated through other means that are less explicit, for instance to account for cases when such deliberate decisions or answers are not readily available?

An answer to this question of how to represent a social network might lie in in a source that is itself less explicit, language. Narratives regarding social events can inform us as to the relationship structure of a group of acquaintances. Likewise, over the duration of a novel a reader deduces, from the setting and the interactions of characters, the nature of relationships between characters. It is suggested that narrative fiction offers a simulation of the social world around us (Mar & Oatley, 2008). Thus, the same aforementioned information we use to predict properties of relationships between individuals in real life can also be applied to fiction. For example, throughout a novel, it is easy to predict the nature and development of friendships and rivalries. Perhaps social networks can also be acquired from and represented implicitly through linguistic sources.

Social information can indeed be extracted from text. Elson, Dames, and McKeown (2011) successfully generated social networks from fictional text by identifying and then analyzing social conversations. They found that social networks can be constructed by simply determining which characters are likely to converse with one another. Similarly, Agarwal, Corvalan, Jensen, and Rambow (2012) were able to build a social network of characters in Alice and Wonderland by analyzing manually annotated social events occurring in the text. These results suggest that information about social relationships is explicitly stated in language, at least within textual conversations and social events.

There is also evidence that language and statistical linguistic frequencies can reveal perceptual information from the world around us. For instance, Louwerse and Zwaan (2009) tested whether language encodes geographical information by correlating statistical linguistic frequencies between cities with the actual physical distances between those cities. Louwerse and Zwaan (2009) further tested the hypothesis by correlating computationally generated semantic relationship values with the longitude and latitude of cities in the US. The semantic associations between cities in a corpus accurately estimated the physical distance between cities. Geographical estimates for fictional cities show a similar effect (Louwerse & Benesh, 2012), supporting the claim that language encodes geographical information. A similar reasoning can perhaps be applied to social relationships. If the physical or psychological distance between individuals is small, their semantic association might be high. In a number of studies we have shown that perceptual and embodied relations are encoded in language (Hutchinson & Louwerse, under review; Louwerse, 2011; Louwerse & Hutchinson, 2012). Perhaps the same is true of social relationships. It might be the case that social relations are also encoded in language, such relations can be established from statistical linguistic patterns.

In the current paper our objective was to determine if social relationships are also encoded implicitly in language, such that computationally generated social networks from character name co-occurrences can approximate manually generated social networks. In the following paper we tested two hypotheses. First, we hypothesized that if individuals are socially related, they will appear together in the text. Second, we hypothesized that if individuals have more social relations, they will appear more in the text.

In three studies we determined if first-order and higher order co-occurrences of pairs of names correlated with an actual social network of characters as generated by humans. To test our two hypotheses, we extracted the semantic relationships between characters in three popular fiction series that varied on the complexity of the social network, *Twilight* (Meyer, 2005, 2006, 2007, 2008), *A Song of Ice and Fire* (Martin, 1996, 1999, 2000, 2005, 2011), and *Harry Potter* (Rowling, 1998, 1999a, 1999b, 2000, 2003, 2005, 2007). The selection of novels was constrained by the following parameters: a. that the novel series had a sufficient and varying number of characters, b. that the novel series was large enough to build an acceptably sized corpus, and c. that the novel series had an available manually generated social network. Readily available manually generated social networks for novel series are limited, therefore the number of series available for use in this study was quite restricted. In addition, we selected series with various levels of complexity of manually generated social networks in order to generalize across social networks.

In the analyses we used both first-order word co-occurrences and higher-order word co-occurrences through Latent Semantic Analysis (LSA). While first order co-occurrences capture the direct frequency with which two names occur together in a text, LSA captures higher-order semantic relations by mapping words into a continuous high dimensional semantic space (Landauer, McNamara, Dennis, & Kintsch, 2007). These first order and higher order networks were then compared with the actual manually generated networks of social relationships between characters in each series.

2. Study 1: Computational Study with Few Characters

In Study 1, we selected the *Twilight* series to determine if we would be able to successfully extract a social network from a text with a small number of character names. We compared computationally generated relationship maps to a simple manually generated relationship map of relationships in *Twilight* to answer to the question whether language statistics encodes social relationships such that individuals who are socially related together, are linguistically discussed together, and individuals who are socially related more, are linguistically discussed more.

To test these hypotheses, four *Twilight* books were converted to one electronic document used for the research purposes described in this study only. The document consisted of a total of 590,520 words and after filtering out frequent stop words, resulted in a final file with 208,100 words and 18,325 paragraphs.

Manually Generated Social Network Maps

We obtained a manually generated social network of the characters in *Twilight* from Muckety LLC (Muckety LLC., 2012b). Muckety is a news corporation that manually generates maps of relationship influence between relevant individuals in a network. They manually specify networks of influence where each node is related to numerous other nodes via specific types of relationships (e.g., friend, enemy, relative). These relationships are validated using a variety of sources, such as government agencies and organizations, news publications, books, organization web sites, and interviews, and are expectedly costly to produce. Muckety generally generates networks representing current political, financial, and educational communities however they have also constructed a social network representing each of the relationships between characters from the *Twilight* series.

Although Muckety provided a manually generated relationship network, edge weights between nodes were not provided. We thus computed edge weights as follows. Considering that between any two individuals there exist approximately four friendship links (Backstrom, Boldi, Rosa, Ugander, & Vigna, 2012), we calculated an exponentially decreasing value representing higher-order relationships up to four degrees away. First order relationships were assigned a value of 1, relationships separated by one friendship link (or degree of separation) were assigned a value of .5, relationships separated by two friendship links were assigned a value of .25, relationships separated by three friendship links were assigned a value of .125, and relationships separated by four friendship links were assigned a value of .0625. To illustrate, two characters sharing a direct relationship (e.g., friends, partners, spouses) would receive +1. Now imagine a third character, sharing a direct relationship with only one of those two characters. That character and the character with whom he/she is not directly related to would receive + .5 because they both share a relationship with a third party but not with each other. This process was repeated until four relationship links were reached (Backstrom et al., 2012).

First-order co-occurrences

We then computationally generated a social network from first-order co-occurrence frequencies of character names. In order to determine the first-order co-occurrences of character names, we computed the co-occurrence of all combinations of the 21 character names in the *Twilight* novels in a five-word window. A window of five words was selected to avoid any issues with data sparsity while still ensuring character names were co-occurring in the text. To avoid any biases with single word and two-word names (*Edward* versus *Edward Cullen*), we selected the names by which each character was most frequently called while keeping the least ambiguous name (e.g., *Edward Cullen* and *Carlisle Cullen* are both be referred to as the homonym *Cullen* in the text, we therefore selected the names *Edward* and *Carlisle*).

These 21x21 frequency combinations were entered in an MDS analysis using the SMACOF algorithm. The SMACOF algorithm minimizes the sum of squares of the error by optimizing the fit to the distances (as opposed to the squared distances) and is thus preferred over ALSCAL, which results in greater error (Young, 1985). We used default criteria for SMACOF, with the maximum iterations = 100, stress convergence = .0001, and the minimum stress value = .0001. Muckety map data were extracted from a *Twilight* map in the way explained above. The *Twilight* Muckety scores for all relations were also entered in an MDS analysis.

First order co-occurrence frequencies were also entered in an MDS analysis and converged in 35 iterations with normalized raw stress = .07. The *Twilight* Muckety scores converged in 21 iterations, with normalized raw stress = .05. For both datasets, the lowest dimensional solution with acceptable stress led to two dimensional solutions. To preserve those two dimensions, we conducted a bidimensional regression analysis to determine the relationship between the Muckety data and the statistical linguistic frequency data. Tobler (1964) and Friedman and Kohler (2003) introduced bidimensional regressions to compute the mapping of any two planes where values of the dependent variable are presented by a point in space, whereby vectors represent intercept and slope. A bidimensional regression for these Muckety and frequency values yielded a strong correlation, $r = .78$, $p < .001$, $n = 21$. This confirms that a social network generated computationally by word frequency acceptably approximates a social network generated manually by humans. The Muckety map values and the co-occurrence estimates are illustrated in Figures 1 and 2. As the two dimensional plots are quite dense, for legibility, the correlation between Muckety map values and co-occurrence values for the first dimension is represented in Figure 1, and the correlation between Muckety map values and co-occurrence values for the second dimension is represented in Figure 2.

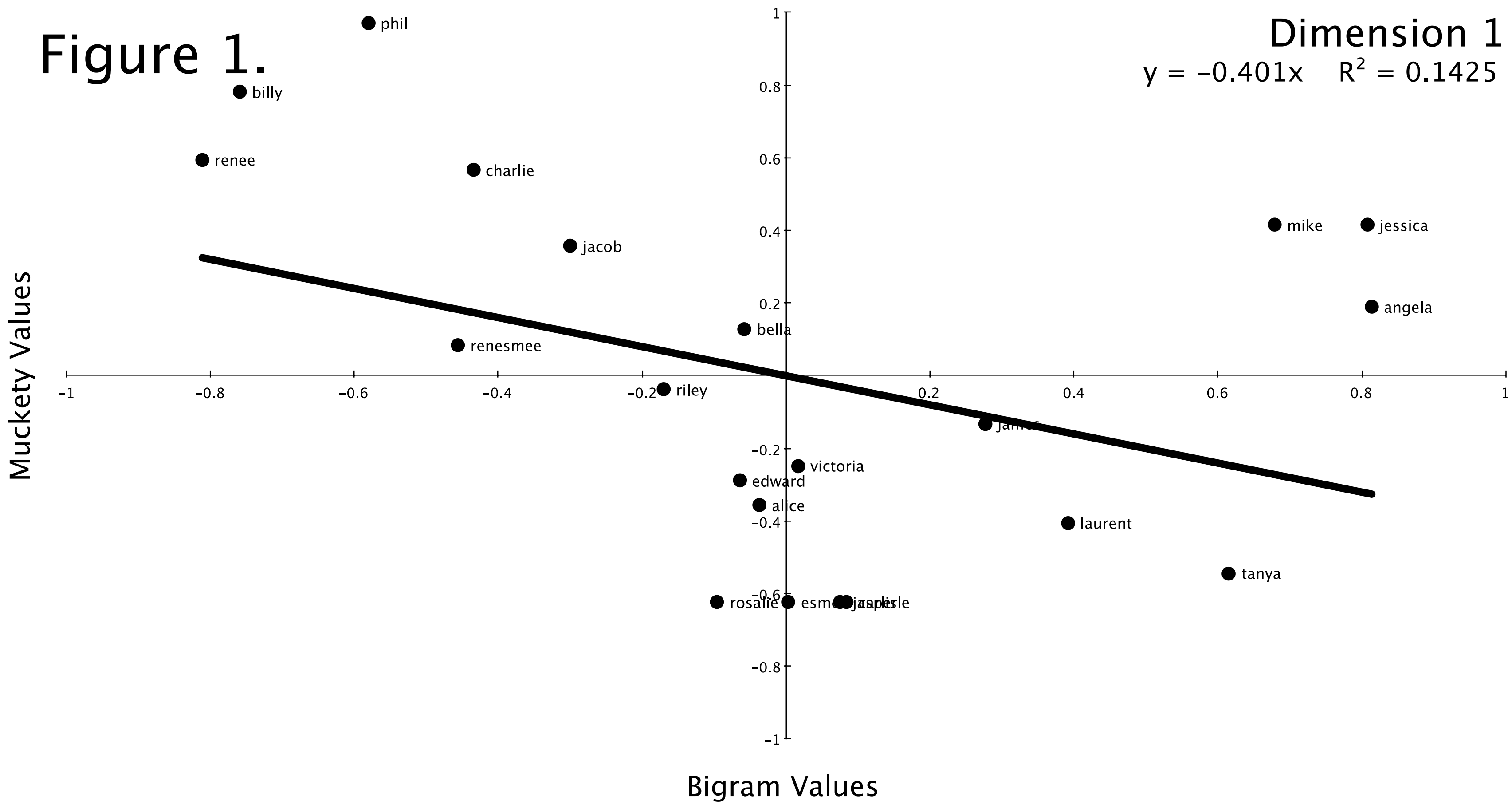

Figure 1. Correlation between bigram values and Muckety map values in *Twilight* for the first dimension representing character prominence.

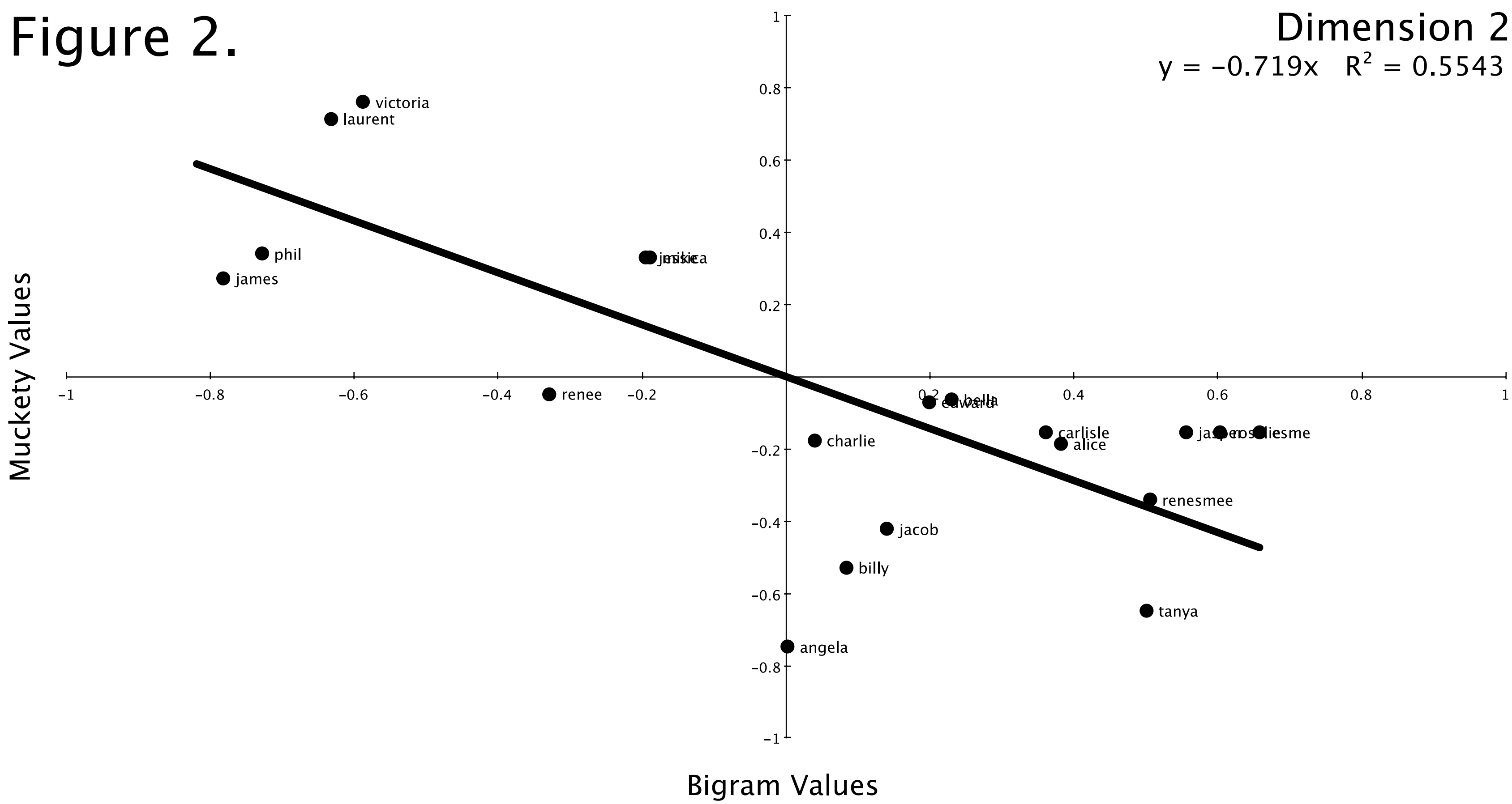

Figure 2. Correlation between bigram values and Muckety map values in *Twilight* for the second dimension representing friends and adversaries.

In order to ensure that the base value of .5 for the edge weights of the Muckety values did not impact the results, we ran the same analysis using a base value of .1. In other words, direct relationships were still assigned a value of 1, but relationships separated by one friendship link were assigned a value of .1, relationships separated by two friendship links were assigned a value of .01, relationships separated by three friendship links were assigned a value of .001, and relationships separated by four friendship links were assigned a value of .0001. The *Twilight* Muckety scores converged in 19 iterations, with normalized raw stress = .08. A bidimensional regression for Muckety and frequency values yielded a similarly strong correlation, $r = .67$, $p < .001$, $n = 21$.

Although first-order frequencies are easy to compute, they also come at a price. Due to sparsity problems, that is, the high probability that characters never co-occur within five numbers of words, they can sometimes give a biased result (Louwerse, 2011). We therefore also used a higher-order co-occurrence algorithm (LSA; Landauer et al., 2007) that does not only compute the probability of two character names occurring within five words but the probability that the neighbors of the neighbors of the neighbors etc., co-occur.

Higher-order co-occurrences

To compute the higher-order computational relationship strength values we employed Latent Semantic Analysis (LSA). More specifically, a first-order process associates stimuli (words) and the contexts they occur in (documents). Stimuli are paired based on their contiguity or co-occurrence. These local associations are next transformed by means of Singular Value Decomposition (SVD) into a small number of dimensions (typically 300) yielding more unified knowledge representations by removing noise.

In the current study the input was the electronic versions of the novels, segmented into paragraphs, from which a large term-document was created. For instance, if there are *m* terms in *n* paragraphs, a matrix of A = $\left(f_{ij} \times G(j) \times L(i, j)\right)_{m \times n}$ was obtained. The value of f_{ij} is a function of the integer that represents the number of times term i appears in document *j*: *L(i; j)* is a local weighting of term *i* in document *j*; and *G(j)* is the global weighting for term *j*. The matrix of *A* has, however, lots of redundant information. Singular Value Decomposition (SVD) reduces this noise by decomposing the matrix *A* into three matrices *A* = $U\Sigma V'$; where *U* is an *m* by *m* and *V* is an *n* by *n* square matrix, with Σ being an *m* by *n* diagonal matrix with singular values on the diagonal. By removing dimensions corresponding to smaller singular values, the representation of each word is reduced as a smaller vector with each word now becomes a weighted vector on 300 dimensions, with only the most important dimensions that correspond to larger singular values being preserved (Landauer et al., 2007). The semantic relationship between characters can then be estimated by taking the cosine between two vectors.

The higher order LSA 21x21 cosine matrix was submitted to MDS, which converged in 80 iterations with normalized raw stress = .10. We again compared the two-dimensional loadings of the Muckety scores and the LSA scores in a bidimensional regression and found a moderate correlation, $r = .48$, $p < .03$, $n = 21$. These findings also suggest that individuals who share a strong social relationship appear together in the text. See Figures 3 and 4 for the correlation between Muckety map values and the LSA estimates.

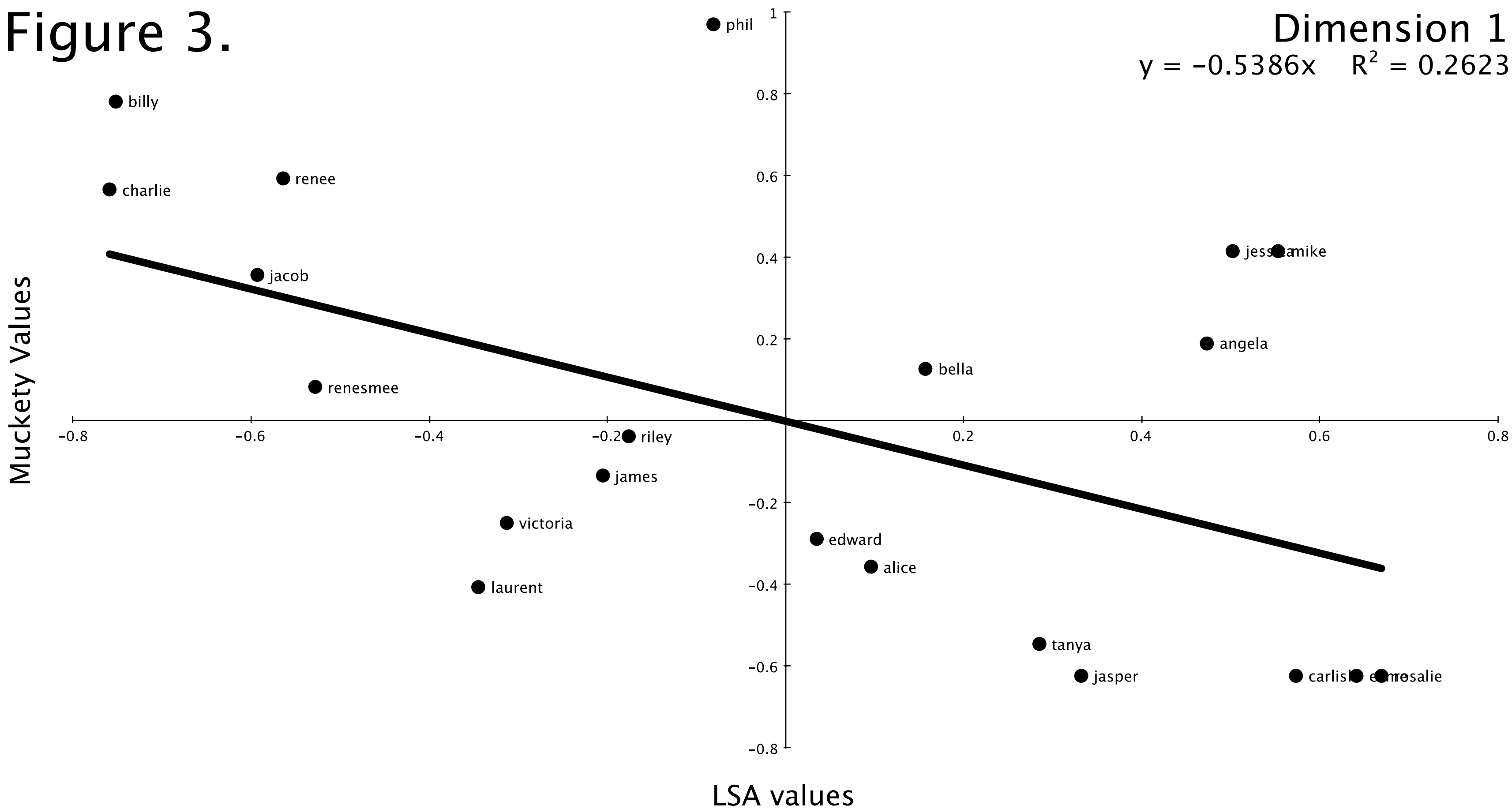

Figure 3. Correlation between LSA values and Muckety map values in *Twilight* for the first dimension representing character prominence.

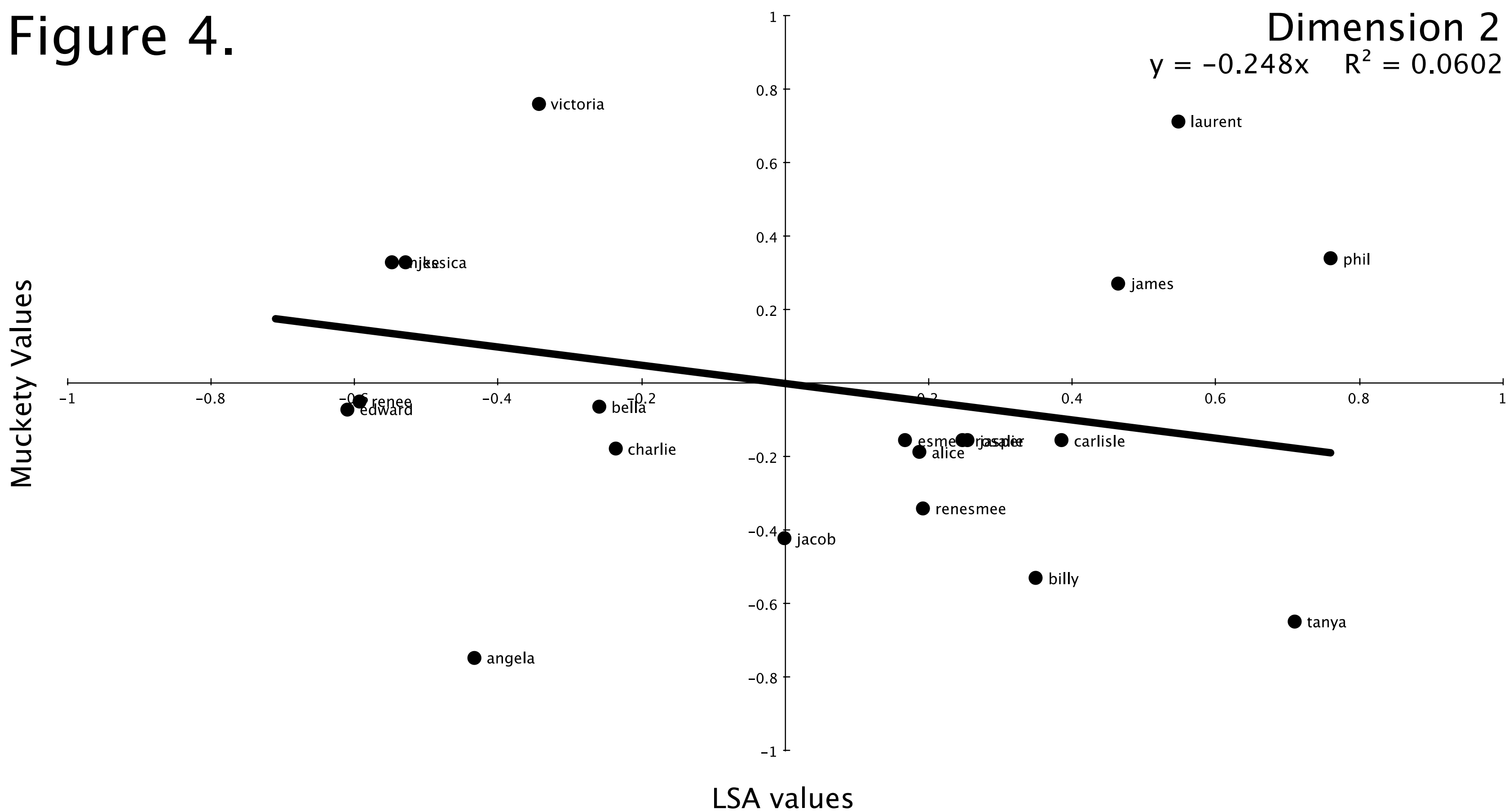

Figure 4. Correlation between LSA values and Muckety map values in *Twilight* for the second dimension representing friends and adversaries.

Again, the question can be raised whether these findings can be explained as a function of the computation of the edge weights. We therefore performed the same analysis after having computed edge weights using a base value of .1. The results with a base value of .1 rather than the earlier base value of .5 did not bias the results, with a bidimensional regression that yielded a similar correlation, $r = .36$, $p = .05$, $n = 21$.

Number of Relationships

Finally, we tested the second hypothesis that characters who were socially more related are linguistically discussed more. First we calculated the frequency of character names in the text and the number of relationships each character had in the Muckety network. We then determined the correlation between name frequency and number of relationships. Name frequency and number of relationships correlated highly, $r = .73$, $p < .001$, $n = 21$, suggesting that individuals who have a large social network appear more in the text.

Discussion

These findings show that social relationships between a small number of characters in a novel are encoded in language. Social networks are inherent in the language itself and can be extracted using both first order and higher order computational methodologies. However, the results might be explained by the relatively small number of characters. Therefore, in Study 2, our objective was to test whether we were able to replicate the results of Study 1 using a more complex character map. In the *Twilight* books, there are very few main characters (only 21) from which to generate a character map. Although these character relations seem to be encoded in language, it could be the case that perhaps this map was not difficult to generate, as there were so few characters.

To demonstrate that language indeed encodes social relations, we aimed to replicate Study 1 using a denser and more complex character relationship map. In addition, to avoid any (unknown) bias to our results from the Muckety network, we used a different source in order to see if our results would generalize across various social networks.

3. Study 2: Computational Study with Many Characters

In Study 2, we analyzed the books *A Song of Ice and Fire* because, unlike Twilight, this series included a large number of characters. Furthermore, we wanted to extend our findings to include a more complex manually generated map from a different source, which was available for this series. Replicating the effects from Study 1 with *A Song of Ice and Fire* would indicate that both small and large social networks can be extracted from text and that these findings are generalizable across different types and sizes of manually generated social network maps.

The map used for the current experiment was manually generated from a collaboratively generated network map posted online (Roseberry, 2012). Whereas the Muckety map had 42 nodes, this complex map had 1,385 nodes. In Study 2, five *A Song of Ice and Fire* books were converted to one electronic document used for the research purposes described in this study only. The document consisted of a total of 1,742,410 words and after filtering frequent stop words, resulted in a final file with 714,098 words and 37,950 paragraphs.

Manually Generated VUE Social Network Map

The collaboratively generated network map posted online (Roseberry, 2012) used the Visual Understanding Environment (VUE), an open source tool used to create visual networks of relationships and information. Using this tool, individuals were able to contribute to and generate a social network for 1,385 characters in the *A Song of Ice and Fire* novels. As with the Muckety maps, each node was related to numerous other nodes via specific types of relationships. Edge weights representing relationship strength were also not provided in this complex map. Edge weights were calculated in the same way as Study 1 until four relationship links were reached.

First-order co-occurrences

The same process as in Study 1 was followed to determine the first-order co-occurrences of character names. We computed the co-occurrence of all combinations of the 80 main character names in a five-word window. These 80 x 80 frequency combinations were entered in an MDS with the maximum iterations = 100, stress convergence = .0001, and the minimum stress value = .0001. Complex map scores for *A Song of Ice and Fire* for all relations were also entered in an MDS analysis.

Co-occurrence frequencies converged in 33 iterations with normalized raw stress = .12. *A Song of Ice and Fire* complex map scores for all relations converged in 49 iterations, with normalized raw stress = .06. Unlike the aforementioned findings, which always resulted in two dimensional solutions representing social relatedness, the lowest dimensional solution with acceptable stress here resulted in a three dimensional solution for the complex map scores and a two dimensional solution for the co-occurrence frequencies. The first dimension of the complex map appeared to account for character proximity, with characters with high values often being found in and around the main city of the story, and characters with lower values being found elsewhere. We therefore selected dimensions 2 and 3 of the complex map that accounted for social relatedness. A bidimensional regression for frequency values and the second and third dimension of the complex VUE map values yielded a correlation of $r = .58$, $p < .001$, $n = 80$. See Figures 5 and 6.

Higher-order co-occurrences

LSA was used to compute higher-order co-occurrences. As before, the 80 x 80 LSA cosine matrix was submitted to an MDS algorithm, which converged in 48 iterations with normalized raw stress = .15. The bidimensional regression for LSA values and the second and third dimension of the complex map values (as above) yielded a moderate (and significant) correlation, $r = .35$, $p = .001$, $n = 80$. The VUE complex map values and the LSA estimates are illustrated in Figures 7 and 8. As both figures show, the correlation between VUE values and LSA loadings are relatively strong. The correlation between VUE values and LSA values for the first dimension is represented in Figure 7, and the correlation between VUE values and LSA values for the second dimension is represented in Figure 8.

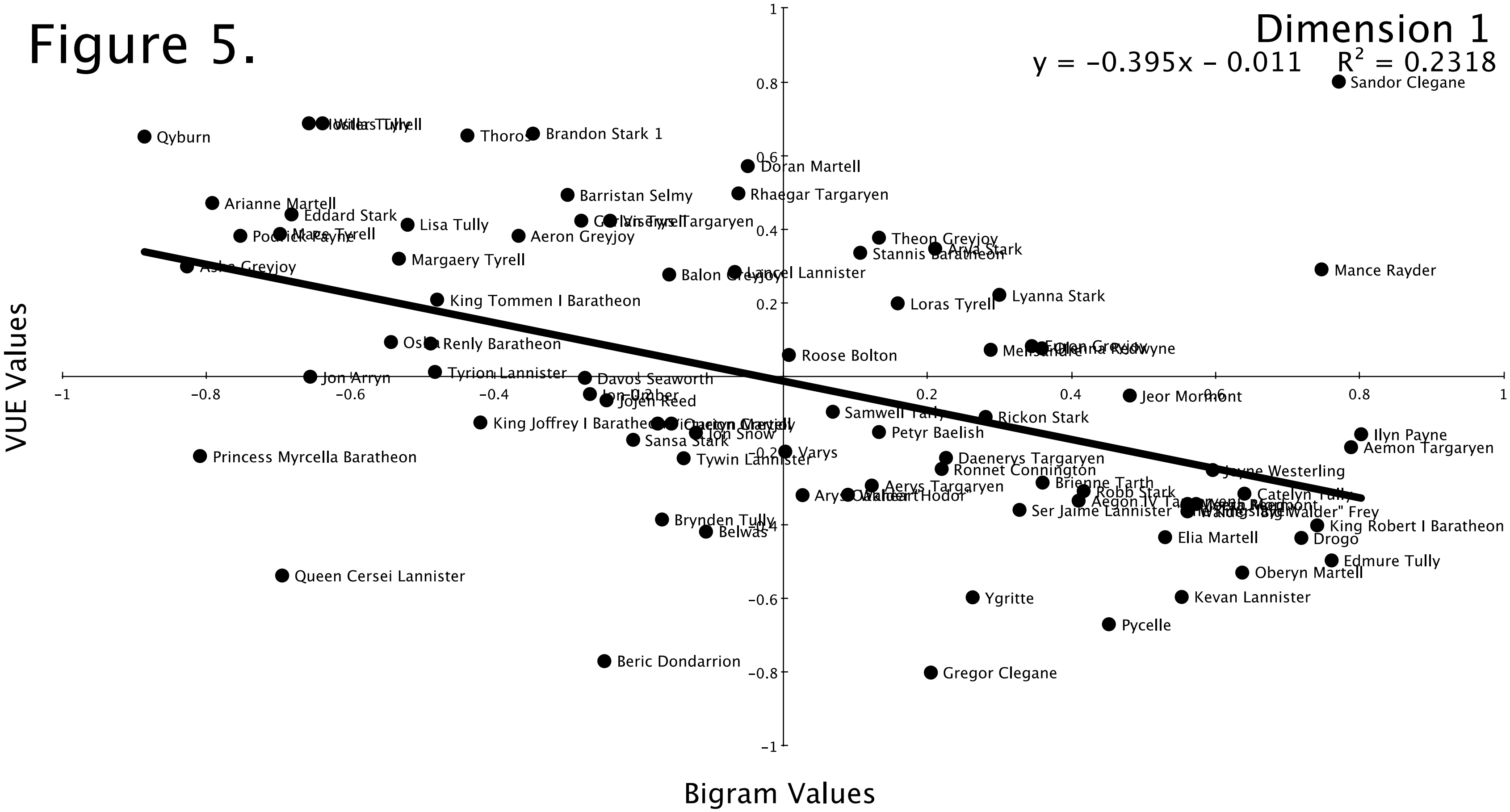

Figure 5. Correlation between bigram values and VUE complex map values in *A Song of Ice and Fire* for the first dimension representing character prominence.

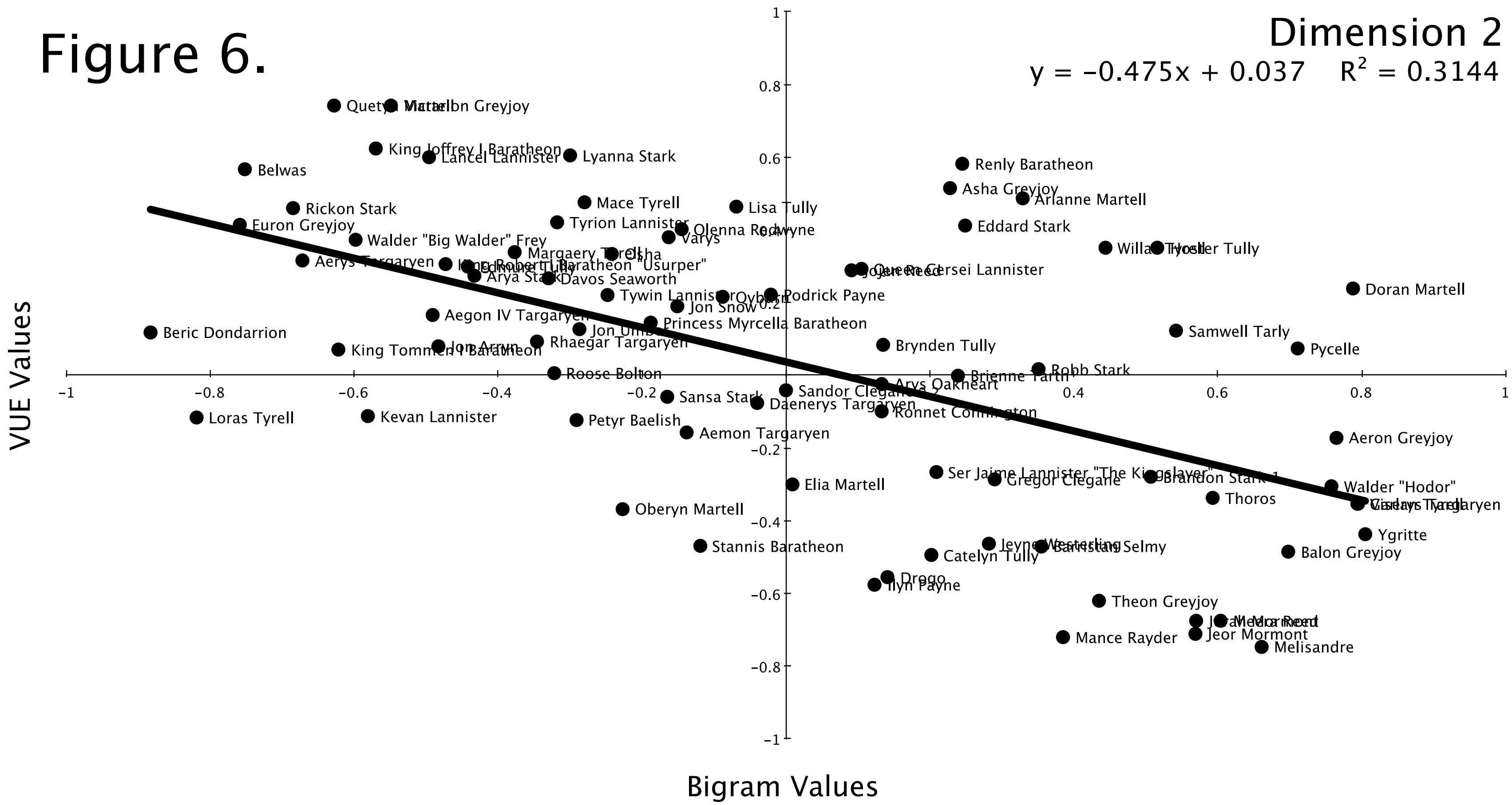

Figure 6. Correlation between bigram values and VUE complex map values in *A Song of Ice and Fire* for the second dimension representing friends and adversaries.

Figure 7.

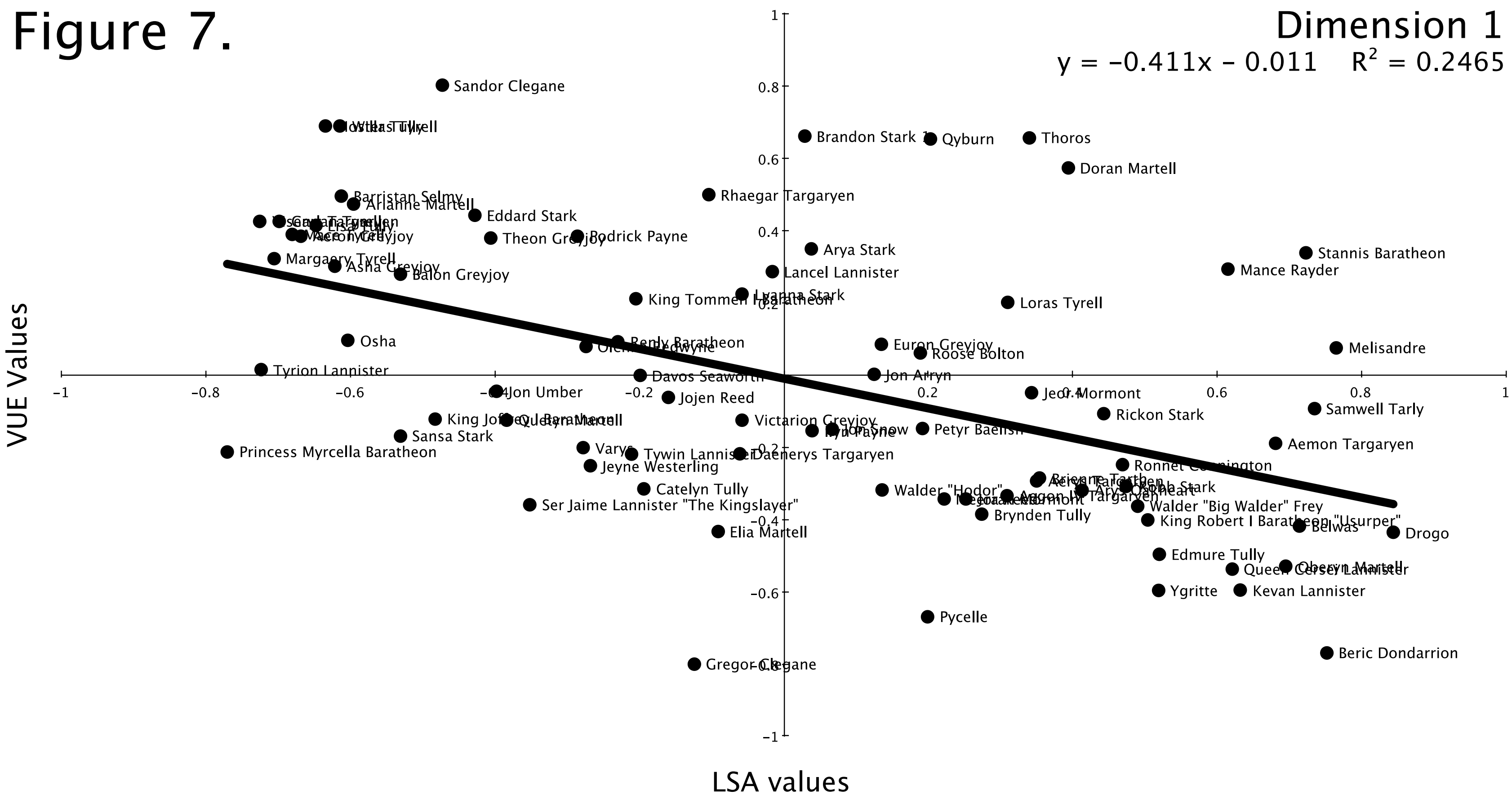

Figure 7. Correlation between LSA values and VUE complex map values in *A Song of Ice and Fire* for the first dimension representing character prominence.

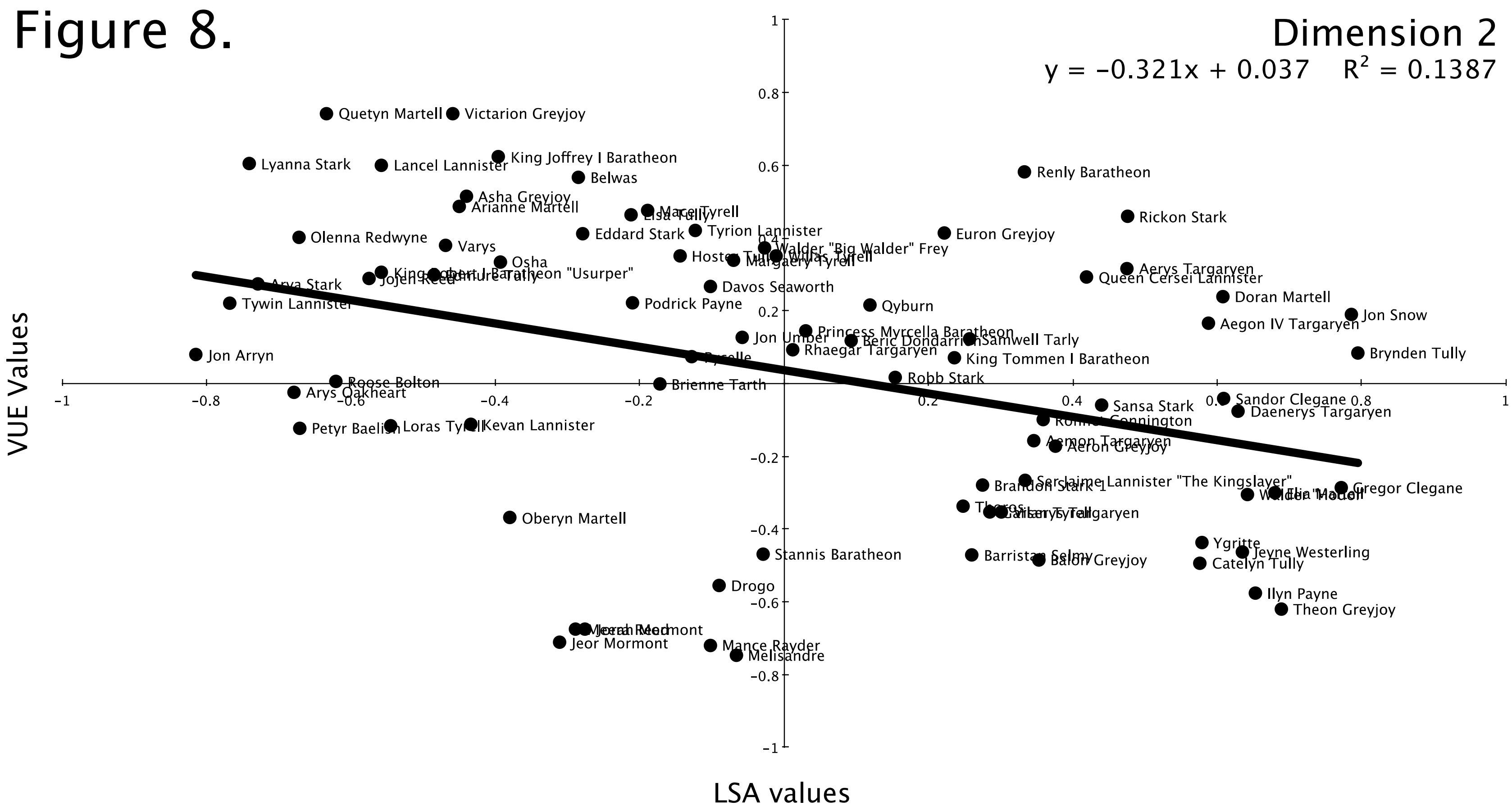

Figure 8. Correlation between LSA values and VUE complex map values in *A Song of Ice and Fire* for the second dimension representing friends and adversaries.

Number of Relationships

To test our second hypothesis that characters who were socially more related appear more in the text we again correlated the frequency of character names in the text and the number of relationships each character had in the VUE network. Name frequency and number of relationships correlated highly, $r = .57$, $p < .001$, $n = 80$.

Discussion

Study 2 confirmed our findings from Study 1 by showing that social networks for a large number of characters are also encoded in language. Not only does this finding hold true for large and small social networks, but it also is generalized across different manually generated social network maps. Both the first-order and the higher-order co-occurrence results again demonstrate that it is possible to extract a social network from language using statistical linguistic frequencies of names of individuals.

4. Study 3a: Computational Study with Moderate Number of Characters

Finally, in Study 3a, we replicated Studies 1 and 2 using *Harry Potter*. This novel was selected because unlike Study 1, *Harry Potter* has greater than 21 characters, while unlike Study 2, a Muckety map was available for the *Harry Potter* series. The seven *Harry Potter* books were converted to one electronic document used for the research purposes described in this study only. The document consisted of a total of 1,277,991, words. The electronic document was then filtered, resulting in a final file with 517,501 words and 21,423 paragraphs.

Manually Generated Social Network Maps

A social network of the characters in *Harry Potter* was obtained from Muckety LLC (Muckety LLC, 2012a). Character relationship values were calculated in the same way as in Studies 1 and 2.

First-order co-occurrences

As in Study 1 and 2, we computed the co-occurrence of all combinations of the 56 character names in the *Harry Potter* novels in a five-word window. Co-occurrence frequencies converged in 10 iterations with normalized raw stress = .16. Similarly, the Muckety scores for all 56 x 56 relations were entered in an MDS analysis and converged in 25 iterations, with normalized raw stress = .13. The bi-dimensional regression for Muckety and co-occurrence values yielded a moderate correlation, $r = .43$, $p < .001$, $n = 56$. See Figures 9 and 10.

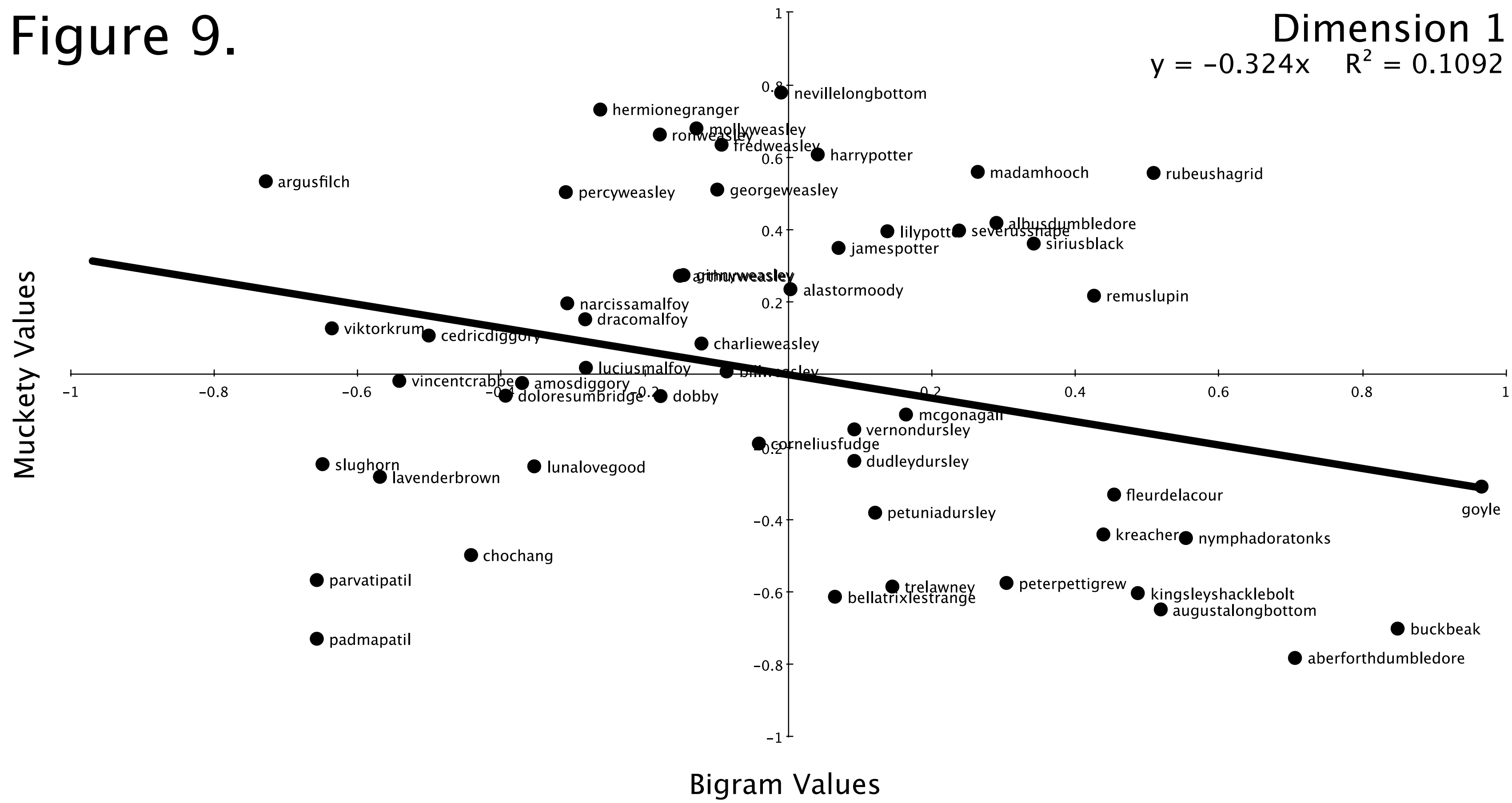

Figure 9. Correlation between bigram values and Muckety values in *Harry Potter* for the first dimension representing character prominence.

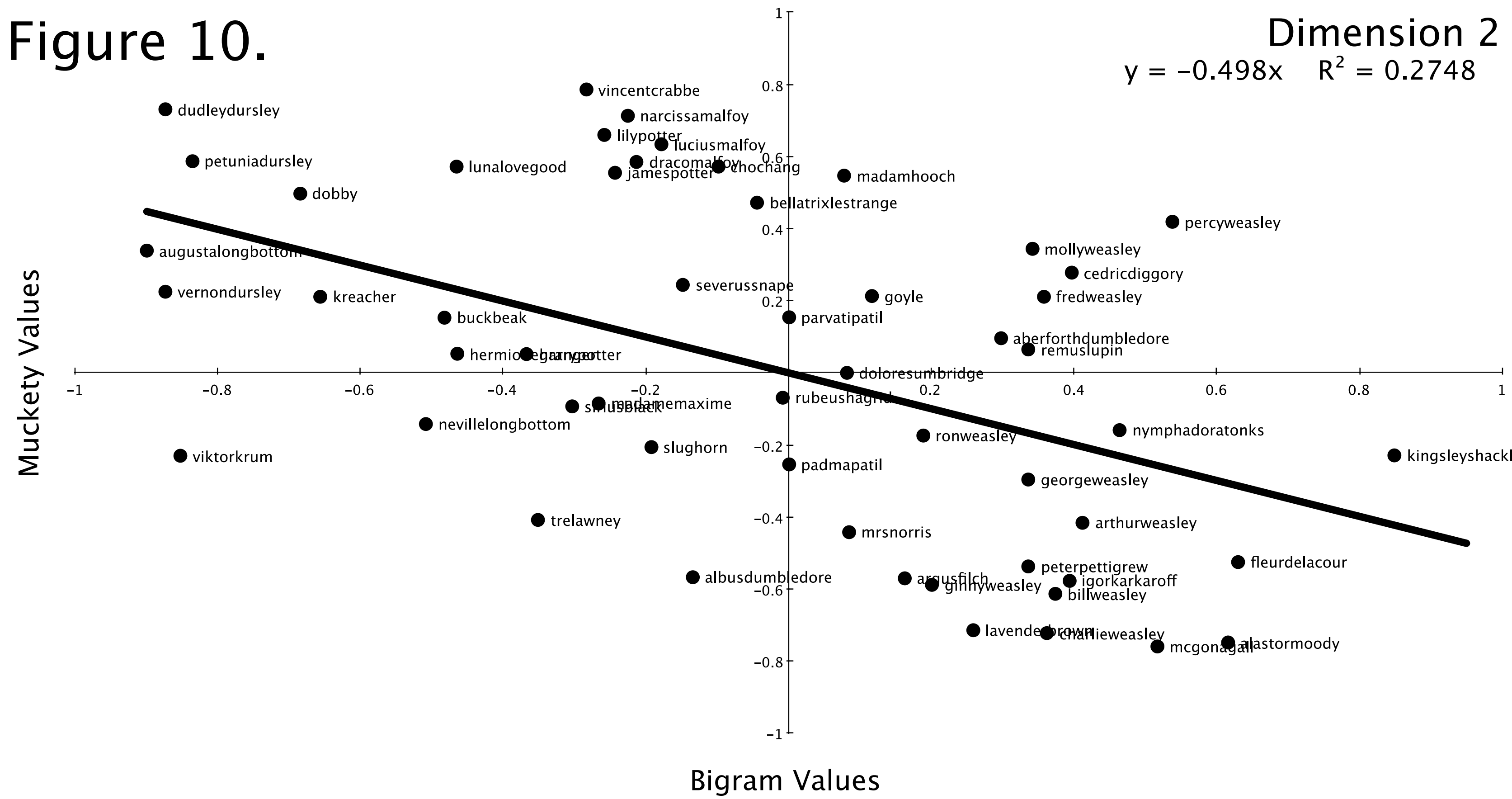

Figure 10. Correlation between bigram values and Muckety values in *Harry Potter* for the second dimension representing friends and adversaries.

Higher-order co-occurrences

To compute the higher-order computational relationship strength values we again employed LSA. When the two-dimensional loadings of the Muckety scores and the LSA scores were compared in a bidimensional regression, somewhat surprisingly, a weak correlation was found, $r = .23$, $p = .08$, $n = 56$. Yet when LSA values were allowed to populate a three dimensional configuration (stress = .07, convergence in 20 iterations), the bi-dimensional regression between Muckety scores and the second and third dimension of the LSA MDS yielded a more moderate (and significant) correlation, $r = .30$, $p = .02$, $n = 56$.

Upon visual inspection of the MDS plot, the first dimension did not explain social relations, but seemed to identify an outlier in the data. The character *Ginny Weasley* had more direct relationships than any of the other characters (except *Harry Potter*), yet the frequency with which *Ginny* occurred in the text was quite low. To illustrate this, the word *Harry* occurred 21,781 times in the text whereas the word *Ginny* only occurred 762 times. After the removal of this outlier we again ran an MDS with two dimensions for both LSA (normalized raw stress = .13, convergence in 12 iterations) and Muckety values (normalized raw stress = .09, convergence in 33 iterations). The bidimensional regression now yielded a strong correlation between LSA values and Muckety values, $r = .76$, $p < .001$, $n = 55$. See Figures 11 and 12.

Figure 11.

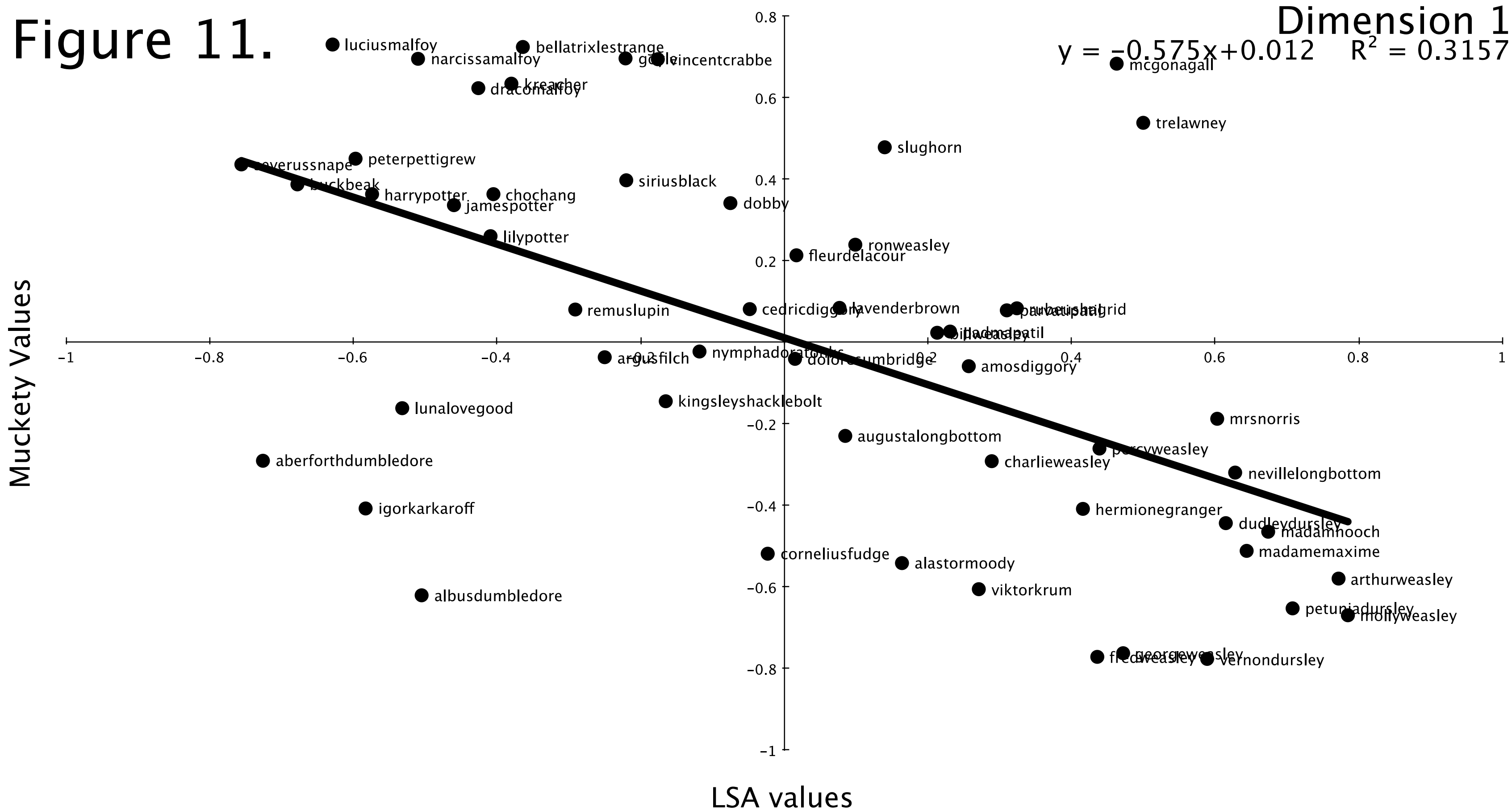

Figure 11. Correlation between LSA values and Muckety values in *Harry Potter* for the first dimension representing character prominence.

Figure 12.

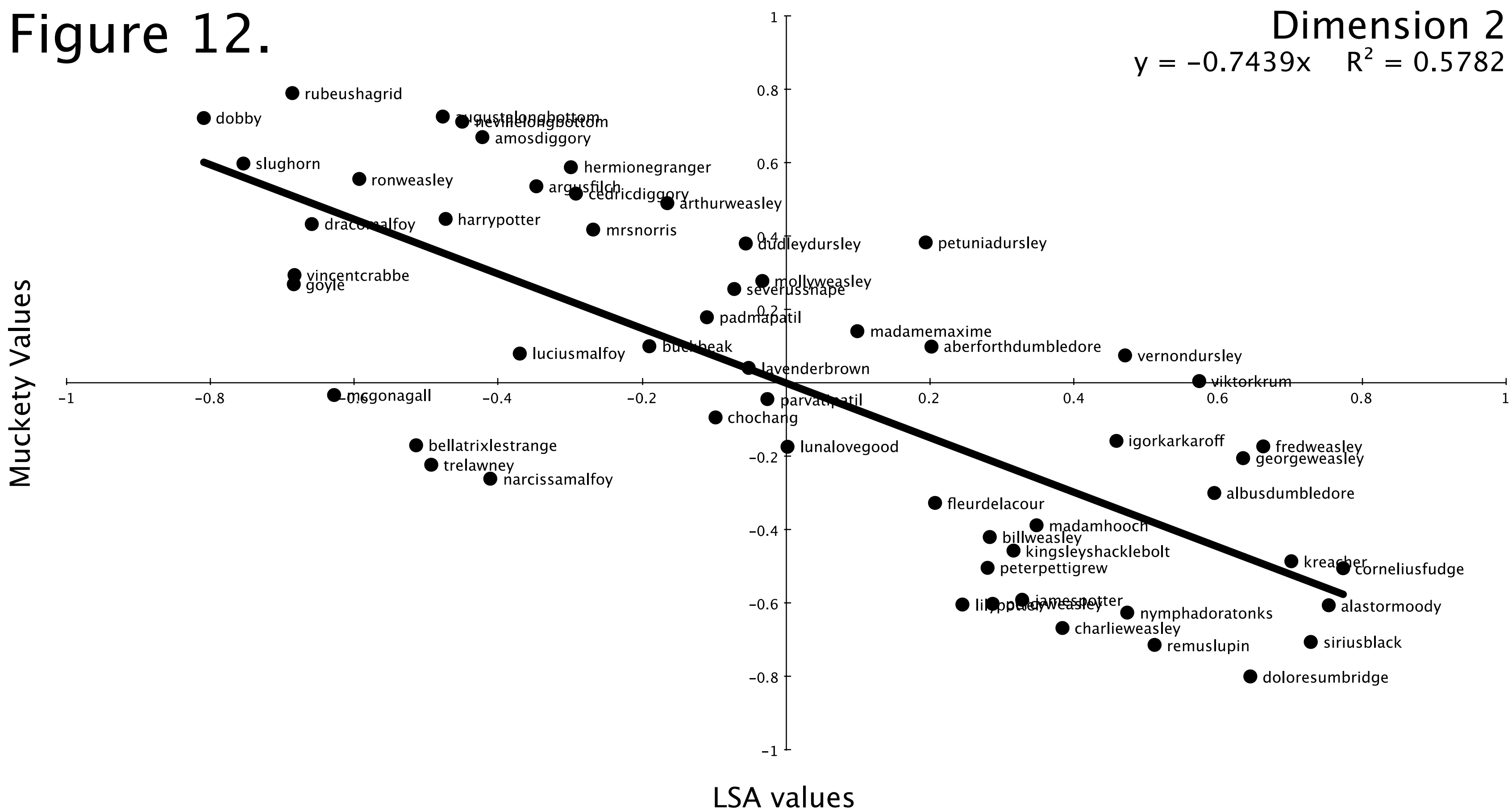

Figure 12. Correlation between LSA values and Muckety values in *Harry Potter* for the second dimension representing friends and adversaries.

In Studies 1 and 2, social networks generated by first order-co-occurrence values performed better than social networks generated by LSA. Initially, this was no different in Study 3, with the LSA network showing a weak correlation to the Muckety map. Only when an outlier was removed did the performance of the LSA network improve. This may have been due in part to the fact that the average ratio of character frequency to number of relations was much higher for this series ($M = .06$, $SD = .2$) than for *Twilight* ($M = .02$, $SD = .04$) or for *A Song of Ice and Fire* ($M = .03$, $SD = .03$). In other words, the removal of one character from the *Harry Potter* LSA network had a greater impact than the removal of a character from either of the other book series, as evidenced by the large increase in r after the removal of a single outlier.

Number of Relationships

Again name frequency and number of relationships correlated highly, $r = .72$, $p < .001$, $n = 56$, suggesting that individuals who have a large social network appear more in the text. By replicating Studies 1 and 2 the above findings of Study 3a also lent support to the conclusion that language encodes social network information.

5. Study 3b: Human Study with Moderate Number of Characters

In Study 3b we collected human data in order to compare our computationally generated maps to social networks generated by experts quite familiar with the character relations in the novels, which also motivated us to use the most popular of these three book series, *Harry Potter*. Although *A Song of Ice and Fire* and the *Twilight* series are well known, for Study 3b we decided to obtain expert networks for the *Harry Potter* series to insure participants were more likely to be very familiar with the characters, as *Harry Potter* is the top selling children's series of books to date.

We investigated whether our findings extended to character maps generated by experts highly familiar with character relations in the novels. Expert readers of *Harry Potter* generated social networks of the main characters in the stories. We compared these expert networks to our first order and higher order computationally generated networks, and to the network obtained from Muckety. This last study determined if computationally generated character maps are also on par with expert human estimates of character relations. So far, we have found evidence that the computational estimates correlate with the estimates from manually generated networks. The question whether expert human ratings compare equally well to the computational estimates is an important question because humans might be using language statistics to build and understand social networks. To address this question, we asked subjects to determine the relationships between characters in the *Harry Potter* series and compared these relationship estimates to both first order and higher order co-occurrences.

Method

Participants

Sixty-six subjects recruited online from Mechanical Turk participated in this study for monetary compensation. All participants were native English speakers. We recruited subjects online in order to increase the likelihood that *Harry Potter* experts would be included in the study. Participants took a 21-item questionnaire consisting of free response questions directly related to the plot of *Harry Potter* (see Appendix). For each of the seven books, there were three questions. These questions were generated and modified by a group of four *Harry Potter* experts who had each read the entire *Harry Potter* series at least two times. These questions allowed us to determine whether participants could be considered *Harry Potter* experts. In addition, participants were informed that knowledge of the *Harry Potter* movies would be insufficient to answer the questions and all participants reported having read all seven books. Fourteen subjects were removed because greater than 20% of their answers to the questionnaire were incorrect.

Stimuli

Each study consisted of close to 500 trials, with each trial including two character names from the *Harry Potter* texts. Character name pairs included all combinations of the 56 main character names assessed in the computational study. Subjects saw a random subset of 500 pairs of characters names.

Procedure

Character name pairs were presented side by side in the center of a computer screen. Participants were asked to indicate the strength of the relationship between the characters on the screen by selecting a value, on a scale of 1 to 6, with 1 being ‘unrelated’ and 6 being ‘closely related’. If participants did not know an answer, they were able to check a box labeled ‘I don’t know’. Once a participant responded, the next trial would commence. Participants were told to answer as accurately as possible. Character name pairs were randomly presented to negate order effects.

Results

The human scores were entered in an MDS analysis, using the same parameters as before. The MDS converged in 23 iterations, with normalized raw stress = .11.

First-order co-occurrences

For our human data, human expert scores for all 56 x 56 relations were entered in an MDS analysis, using the same parameters as for the previous data. The bi-dimensional regression for human and co-occurrence values yielded a correlation of $r = .26$, $p = .05$, $n = 56$.

Unlike before, the correlations between human relationship strength values and computational values were much weaker than the correlations between human generated maps and the same computational values. However, keep in mind that the human generated maps (Muckety and VUE maps) were created collaboratively over time with reference to the novels. In contrast, the human relationship strength estimates here were made by individual subjects over the duration of one hour without ready access to the *Harry Potter* novels. Therefore, we sought to establish whether it might be the case that the relationships of important characters might be better estimated by humans in an experimental session than the relationships of unimportant characters. To do so, we classified important characters as those having the top 25% of number of relationships. The same analyses were then conducted with these fourteen characters. The bidimensional regression for human and co-occurrence values yielded a much stronger correlation of $r = .45$, $p = .05$, $n = 14$. See Figures 13 and 14.

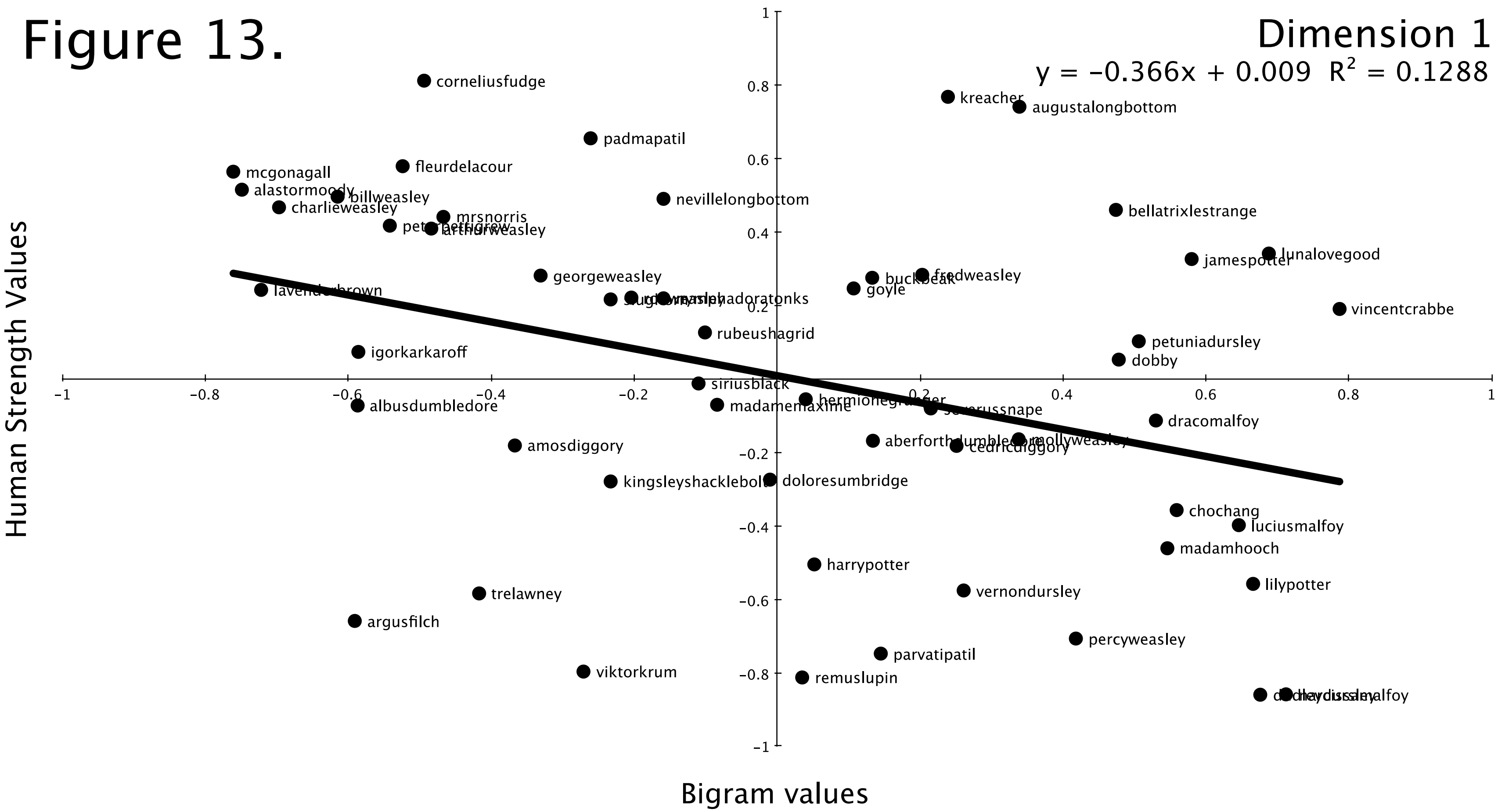

Figure 13. Correlation between bigram values and human map values in *Harry Potter* for the first dimension representing character prominence.

Figure 14.

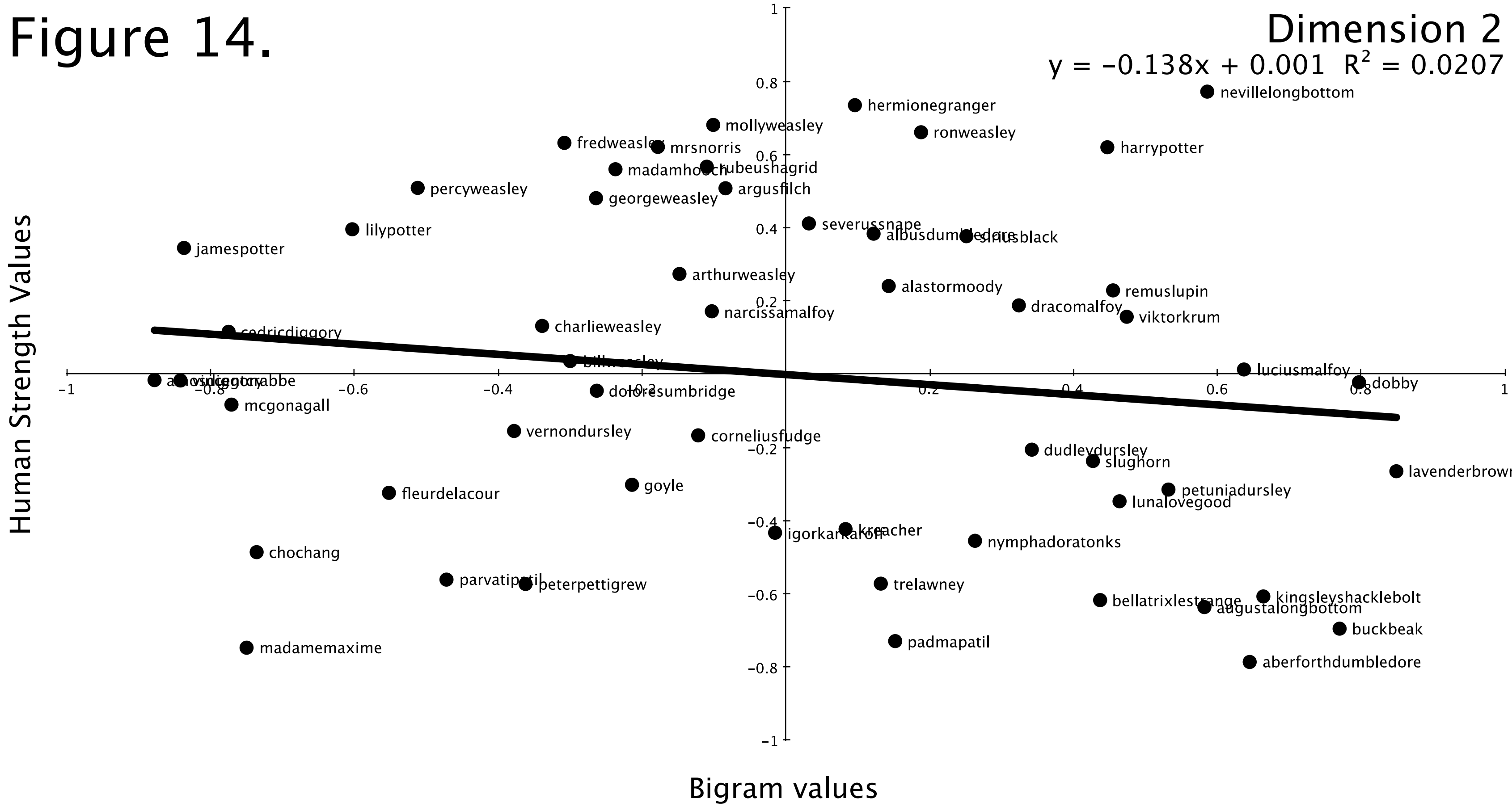

Figure 14. Correlation between bigram values and human map values in *Harry Potter* for the second dimension representing friends and adversaries.

Higher-order co-occurrences

We used the same human expert loadings and LSA cosine values as before to compare human scores and the LSA MDS. For higher-order computational relationship strength values, the 56 x 56 LSA cosine matrix converged in 11 iterations with normalized raw stress = .15. When the two-dimensional loadings of the human scores and the LSA scores were compared, a moderate correlation was found, $r = .34$, $p = .01$, $n = 56$. When LSA values populated a three dimensional configuration as in Study 3a (stress = .07, convergence in 20 iterations), the bi-dimensional regression between human scores and the second and third dimension of the LSA MDS yielded a stronger correlation, $r = .39$, $p < .01$, $n = 56$.

Because the outlier *Ginny Weasley* accounted for the first dimension of the LSA MDS, we again ran an MDS after removing *Ginny Weasley* for both LSA (normalized raw stress = .13, convergence in 12 iterations) and human values (normalized raw stress = .10, convergence in 19 iterations). The bidimensional regression yielded a similar correlation between the first two dimensions of LSA values and human values, $r = .32$, $p < .001$, $n = 55$ and the results above excluding the first dimension.

Again, to see if important characters might be better estimated by humans in an experimental session than the relationships of unimportant characters, when analyzing only the characters with the top 25% of number of relationships, a stronger correlation was also found for human and LSA values, $r = .63$, $p < .02$, $n = 14$. See Figures 15 and 16.

Figure 15.

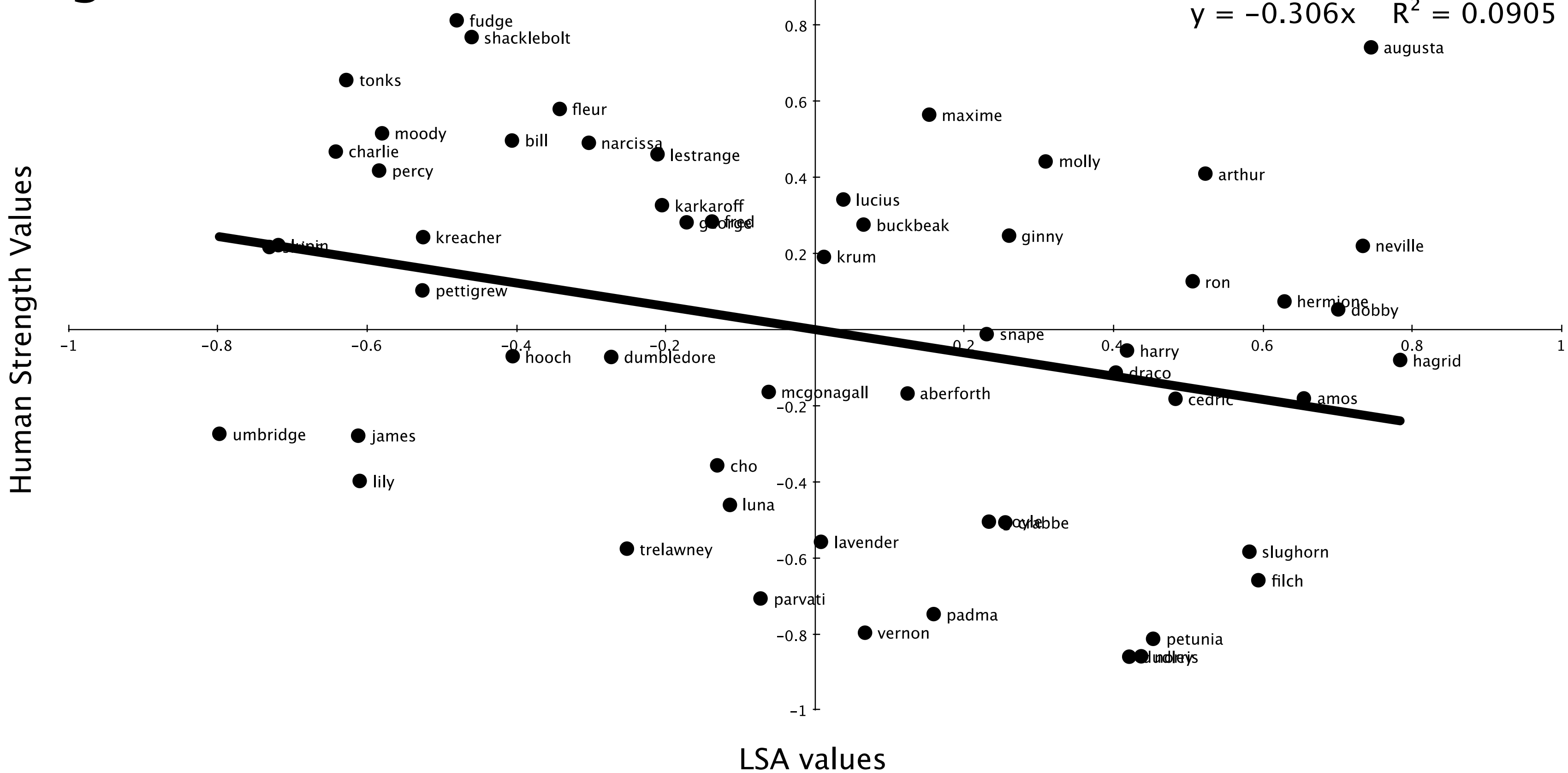

Figure 15. Correlation between LSA values and human map values in *Harry Potter* for the first dimension representing character prominence.

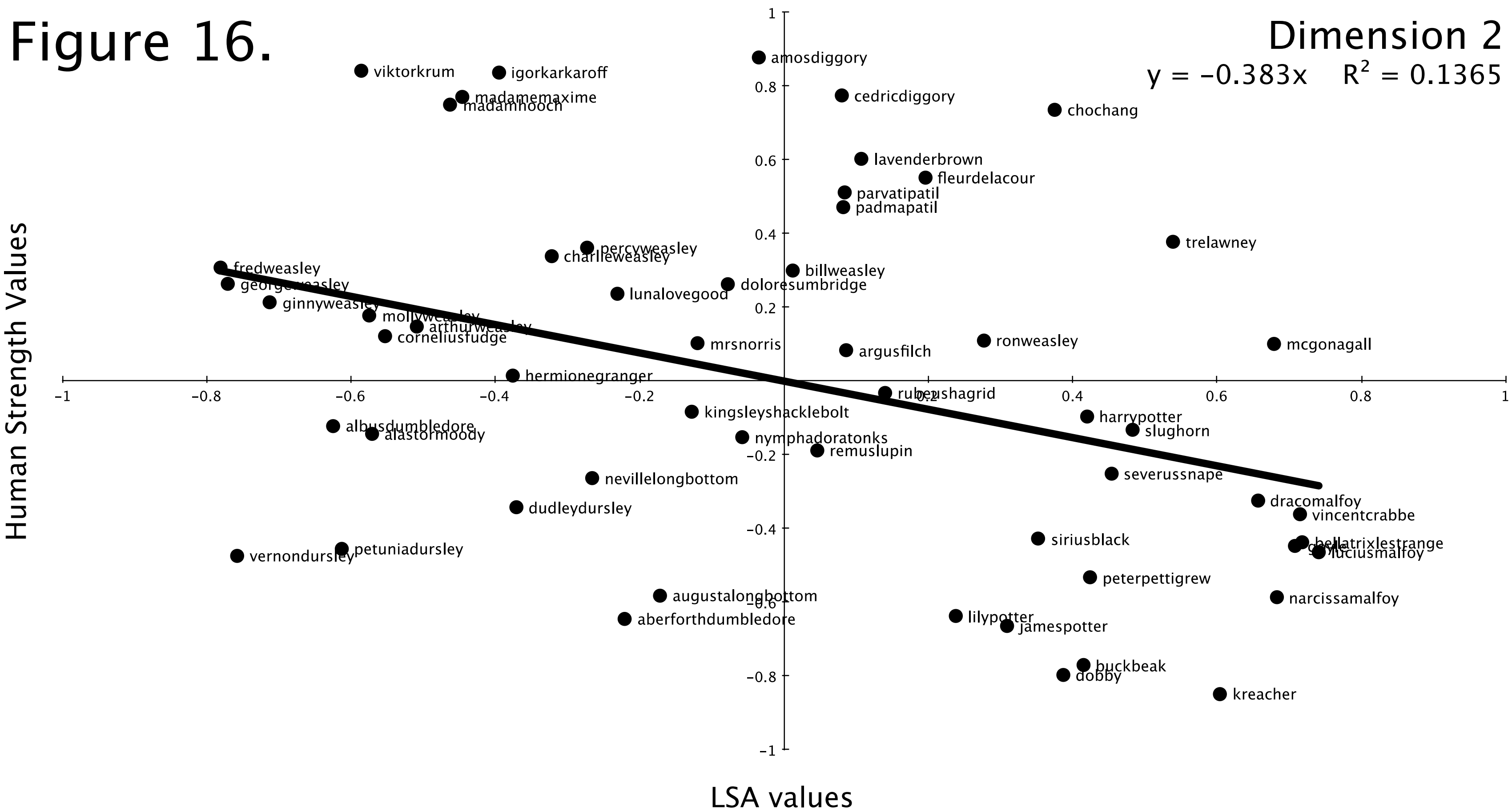

Figure 16. Correlation between LSA values and human map values in *Harry Potter* for the second dimension representing friends and adversaries.

Number of Relationships

To explore our second hypothesis, in Study 3b we calculated number of relationships as the count of those relationships that were strong (5-6 on a 1-6 scale). Supporting this hypothesis again, name frequency and number of relationships correlated highly, $r = .69$, $p < .01$, $n = 56$.

Discussion

Studies 3a and 3b demonstrated further that social networks are inherent in language itself. Study 3a nicely replicated results from Studies 1 and 2, whereby LSA and first-order co-occurrences were able to approximate a manually generated social network map obtained from Muckety LLC. In Study 3b we further demonstrated that LSA and first-order approximations of social networks adequately correlated with expert human estimates of character relations. So far, we have found evidence that the computational estimates correlate with the estimates from manually generated networks. The question whether expert human ratings compare equally well to the computational estimates is an important question because humans might be using language statistics to build and understand social networks

6. General Discussion

The current studies aimed to determine if language encodes social relationships. The reported results suggest computationally derived character pair values can explain relationship networks generated by humans for three fictional novel series, *Twilight*, *A Song of Ice and Fire*, and *Harry Potter*. These findings demonstrate that individuals who are socially related together are linguistically discussed together. For the first set of analyses we used first-order co-occurrences that yielded acceptable bidimensional regression coefficients. A set of higher-order co-occurrence (LSA) analyses also yielded reliably high bidimensional regression coefficients. Finally, we found evidence that those individuals with more social relationships appear more frequently in the text, as there were strong positive correlations between number of relationships and unigram frequency of character names.

Even though narrative fiction offers a simulation of the social world around us (Mar & Oatley, 2008), the main conclusion of this study can be extended to the non-fictional world. The social networks here were created from corpora of fictional novels, but it might be the case that social networks about actual individuals can also be extracted from language, given the right corpus. We have already demonstrated this for geographical estimates for cities in the United States using newspapers (Louwerse & Zwaan, 2009), and geographical estimates for cities in the fictional Middle Earth using Lord of the Rings (Louwerse & Benesh, 2012). We therefore expect that this method for extracting social networks from fictional novels can be extended to non-fictional texts. For instance, by using newspaper articles, social relations among political and financial leaders can be determined. By using blogs and tweets, social networks of individuals in these texts might also be estimated. However, non-fictional corpora might include examples of relationships that are not necessarily social, but perhaps functional, or hierarchical (Fischer, 1982). Whereas in a novel characters are involved in the same events and environments, news articles might refer to individuals who share some attribute but who are not necessarily socially related.

In addition, an important conclusion for the cognitive sciences is that language encodes perceptual and physical relations in the world around us, such as social relations.

However, if social networks are encoded in language, the question arises whether humans use these cues when understanding social networks in the real world. Rather, do human beings obtain social information from language? As demonstrated here, we do seem garner some social network information from language. Social relations are established based on aspects such as proximity, familiarity, physical attributes, emotional reactions, shared experiences, duration, reciprocation, and similar attitudes, among a multitude of other factors (Byrne, 1971; Crandall et al., 2010; Ebbesen et al., 1976; Reis et al., 2011). Thus, these same factors we use to predict properties of relationships between individuals seem to be mirrored in language such that approximated social networks are represented implicitly through the co-occurrences of character names alone. In other words, even without the explicit consideration of even one factor relevant to establishing and creating relationships, computational linguistic techniques are able to generate approximate social networks of a set of characters in a corpus. And although much information is extracted from name co-occurrences alone, each of these additional factors might also be taken into consideration when understanding or building a social network. As demonstrated in the human study, it might be the case that humans rely on linguistic cues to generate networks for those salient characters but not for those who are ancillary, and further research is needed to determine to what extent humans rely on linguistic cues to understand and generate social networks. In previous work we have shown language implicitly encodes geographical information (Louwerse & Benesh, 2012; Louwerse, Hutchinson, & Cai, 2012; Louwerse & Zwaan, 2009) and other types of perceptual information (Hutchinson & Louwerse, under review; Louwerse, 2011; Louwerse & Hutchinson, 2012). The current study shows that this can be extended to social information. Language has evolved such that statistical linguistic frequencies can capture the social relationships in the world around us. Social relations are encoded in language, such that humans can use language statistics to build and understand social networks.

References

Agarwal, A., Corvalan, A., Jensen, J., & Rambow, O. (2012). Social Network Analysis of Alice in Wonderland. In A. Kazantseva, D. Elson, & S. Szpakowicz (Eds.), *Proceedings of the Workshop on Computational Linguistics for Literature* (pp. 88-97). Montreal, Canada: The Association for Computational Linguistics.

Backstrom, L., Boldi, P., Rosa., Ugander., & Vigna, S. (2011). *Four degrees of separation*. Retrieved from http://arxiv.org/abs/1111.4570

Byrne, D. (1971). *The attraction paradigm*. New York, NY: Academic Press.

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., & Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, *107*, 22436–22441.

Ebbesen, E. B., Kjos, G. L., & Konecni, V. J. (1976). Spatial ecology: Its effects on the choice of friends and enemies. *Journal of Experimental Social Psychology*, *12*, 505–518.

Elson, D. K., Dames, N., & McKeown, K. R. (2010). Extracting social networks from literary fiction. In J. Hajic, S. Carberry, & S. Clark (Eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 138–147). Uppsala, Sweden: Association for Computational Linguistics.

Farina, A., Wheeler, D. S., & Mehta, S. (1991). The impact of an unpleasant and demeaning social interaction. *Journal of Social and Clinical Psychology*, *10*, 351–371.

Feeney, J. A., & Noller, P. (2011). Attachment style and romantic love: Relationship dissolution. *Australian Journal of Psychology*, *44*, 69–74.

Festinger, L., Schachter, S., & Black, K.W. (1950) *Social pressures in informal groups: A study of human factors in housing*. New York, NY: Harper

Fischer, C. S. (1982). *To dwell among friends: Personal networks in town and city*. Chicago, IL: University of Chicago Press.

Friedman, A., & Kohler, B. (2003). Bidimensional regression: A method for assessing the configural similarity of cognitive maps and other two-dimensional data. *Psychological Methods, 8*, 468–491.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*, 1360–1380.

Hatfield, E., Traupmann, J., & Walster, G. W. (1978). Equity and extramarital sexuality. *Archives of Sexual Behavior, 7,* 127-141. Reprinted in M. Cook & G. Wilson (Eds.). (1979). *Love and attraction: An international conference* (pp. 309-323). Oxford, England: Pergamon Press.

Hutchinson, S., & Louwerse, M. M. (2012). *Statistical linguistic context and embodiment predict metaphor processing but participant gender determines how much.* Manuscript submitted for publication.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Li, W., Moallem, I., Paller, K. A., & Gottfried, J. A. (2007). Subliminal smells can guide social preferences. *Psychological Science, 18,* 1044-1049.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *TopiCS in Cognitive Science, 3,* 273-302.

Louwerse, M.M., & Benesh, N. (2012). Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. *Cognitive Science, 36,* 1556-1569.

Louwerse, M. M., & Hutchinson, S. (2012). Neurological Evidence Linguistic Processes Precede Perceptual Simulation in Conceptual Processing. *Frontiers in Psychology, 3.* 385.

Louwerse, M. M., Hutchinson, S., & Cai, Z. (2012). The Chinese route argument: Language predicts longitude and latitude of locations. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 695-700). Austin, TX: Cognitive Science Society.

Louwerse, M. M., & Zwaan, R.A. (2009). Language encodes geographical information. *Cognitive Science*, *33*, 51-73.

Mar, R. A., & Oatley, K. (2008). The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science*, *3*, 173–192.

Martin, G. R. R. (1996). *A game of thrones.* New York, NY: Bantam Books.

Martin, G. R. R. (1999). *A clash of kings.* New York, NY: Bantam Books.

Martin, G. R. R. (2000). *A storm of swords.* New York, NY: Bantam Books.

Martin, G. R. R. (2005). *A feast for crows.* New York, NY: Bantam Books.

Martin, G. R. R. (2011). *A dance with dragons.* New York, NY: Bantam Books.

Meyer, S. (2005). *Twilight.* New York, NY: Little Brown and Company.

Meyer, S. (2006). *New moon.* New York, NY: Little Brown and Company.

Meyer, S. (2007). *Eclipse.* New York, NY: Little Brown and Company.

Meyer, S. (2008). *Breaking dawn.* New York, NY: Little Brown and Company.

Muckety LLC. (2012a). *Harry Potter Series* [Graphical Interactive Relationship Influence Map]. Retrieved from http://www.muckety.com/Harry-Potter-series/5017817.muckety

Muckety LLC. (2012b). *Twilight Series* [Graphical Interactive Relationship Influence Map]. Retrieved from http://www.muckety.com/The-Twilight-Saga/5062899.muckety

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.

Reis, H. T., Maniaci, M. R., Caprariello, P. A., Eastwick, P. W., & Finkel, E. J. (2011). Familiarity does indeed promote attraction in live interaction. *Journal of Personality and Social Psychology*, *101*, 557–570.

Roseberry, G. (2012). *A Song of Ice and Fire Series* [Visual Understanding Environment Relationship Map]. Retrieved from http://dl.dropbox.com/u/10153810/basic%20v1.vue

Rowling, J. K. (1998). *Harry Potter and the sorcerer's stone.* New York, NY: Scholastic Books.

Rowling, J. K. (1999a). *Harry Potter and the chamber of secrets.* New York, NY: Scholastic Books.

Rowling, J. K. (1999b). *Harry Potter and the prisoner of Azkaban.* New York, NY: Scholastic Books.

Rowling, J. K. (2000). *Harry Potter and the goblet of fire.* New York, NY: Scholastic Books.

Rowling, J. K. (2003). *Harry Potter and the order of the phoenix.* New York, NY: Scholastic Books.

Rowling, J. K. (2005). *Harry Potter and the half blood Prince.* New York, NY: Scholastic Books.

Rowling, J. K. (2007). *Harry Potter and the deathly hallows.* New York, NY: Scholastic Books.

Scott, J. (1988). Social network analysis. *Sociology*, *22*, 109–127.

Sprecher, S., & Hendrick, S.S. (2004). Self-disclosure in intimate relationships: Associations with individual and relationship characteristics over time. *Journal of Social & Clinical Psychology, 23*, 857-877

Takhteyev, Y., Gruzd, A., & Wellman, B. (2011). Geography of twitter networks. *Social Networks*, *34*, 73-81.

Tobler, W. R. (1964). Bidimensional regression. *Geographical Analysis, 26,* 187–212.

Touhey. J . C. (1972) Comparison of two dimensions of attitude similarity on heterosexual attraction. *Journal of Personality and Social Psychology, 23,* 8-10.

Young, F.W. (1985) Multidimensional scaling. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, (Vol. 5, pp. 649-659). New York, NY: Wiley.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Applied Social Psychology, 9,* 1-27.

Zajonc, R.B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science, 10*, 224-228.

Appendix A

Questionnaire designed to test knowledge of the Harry Potter books

BOOK 1: Harry Potter and the Sorcerers' Stone

What is the name of the goblin who escorted Harry to his vault?

Where did Harry first meet Draco?

What piece does Ron play on the giant chess board?

BOOK 2: Harry Potter and the Chamber of Secrets

In book two, where was Harry when he first heard the snake voice that nobody else could hear?

Why does Ron think that Filch is so bitter?

At Hogwarts, where is the entrance to the Chamber of Secrets?

BOOK 3: Harry Potter and the Prisoner of Azkaban

How did Harry get to Diagon Alley after he fled the Dursley's?

Who are the authors of the Maurader's map?

How did Hermione manage to take several classes at the same time during her third year at Hogwarts?

BOOK 4: Harry Potter and the Goblet of Fire

Harry survives the first task with the help of what?

Who does Harry want to take to the Yule Ball?

How does Rita Skeeter get her scoops?

BOOK 5: Harry Potter and the Order of the Phoenix

Who sent Dementors to attack Harry's house?

A group of students form a secret study group to practice Defense Against the Dark Arts. What did Cho Chang suggest they call it?

Name a professor whom Umbridge had dismissed from Hogwarts.

BOOK 6: Harry Potter and the Half Blood Prince

What did Harry use that helped him obtain the memory from Slughorn?

Whom did Harry choose as the Gryffindor keeper?

Whom did Harry take to Slughorn's Christmas Party?

BOOK 7: Harry Potter and the Deathly Hallows

Who does Voldemort initially borrow a wand from in hopes of defeating Harry?

Where do Harry, Ron, and Hermione venture to find Hufflepuff's cup?

Where is the gravesite of Lily and James Potter?

Appendix B

IRB Review and Approval

THE UNIVERSITY OF MEMPHIS

Institutional Review Board

To:	Max Louwerse Psychology
From:	Chair or Designee, Institutional Review Board For the Protection of Human Subjects irb@memphis.edu
Subject:	Cognitive Maps of Cities (#2388, E07-127)
Approval Date:	October 3,2012

This is to notify you that the Institutional Review Board has designated the above referenced protocol as exempt from the full federal regulations. This project was reviewed in accordance with all applicable statuses and regulations as well as ethical principles.

When the project is finished or terminated, please submit a Human Subjects Research Completion Form (COMP) to the Board via e-mail at irbforms@memphis.edu. This form can be obtained on our website at http://www.memphis.edu/irb/forms.php.

Approval for this protocol does not expire. However, any change to the protocol must be reviewed and approved by the board prior to implementing the change.